A Dissertation Proposal

On

Factors Affecting The Performance Of Data Mining Tools

Submitted To

Lovely Professional University

In partial fulfillment of the requirement for the award of degree of

MASTER OF PHILOSPHY (M.Phil)

In

COMPUTER SCIENCE

Submitted by: **Supervised By:**

Balrajpreet Kaur **Dr. Anil Sharma**

Reg. No. 11512454

SCHOOL OF COMPUTER APPLICATION

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA (PUNJAB)

CERTIFICATE OF THE SUPERVISOR

This is to certify that the work Factors affecting the performance of data mining tools is a section of research work done by Balrajpreet Kaur under my guidance and supervision for the degree of Master of Philosophy (M.Phil.) in Computer Science of Lovely Professional University, Phagwara, Punjab, India. To the best of my knowledge , the present work is the result of her original analysis and study. No part of the project report has ever been submitted for any degree or diploma. The dissertation is apt for the submission for partial fulfillment of the conditions for the award of M.Phil in Computer Science.

Date: Signature of Supervisor

Dr. Anil Sharma

DECLARATION

I hereby admit that the dissertation proposal entitled, "**Factors Affecting The Performance Of Data Mining Tools**" is an authentic record of my own original work carried out for the award of degree of M.Phil (Computer Science) and all ideas and references have been duly acknowledged. The matter presented in the dissertation has not been submitted in part or full to any other university or institute for the award of any degree.

Date: **Signature of the student**

Balrajpreet Kaur
Reg. No: 11512454

ACKNOWLEGEMENT

The report has been written with the kind guidance and support of my Supervisor. The satisfaction and happiness that accompany the successful completion of any task would be incomplete without mentioning the names of people who made it possible, whose constant guidance and encouragement crowns all efforts with our success.

I would like to express my deep gratitude and thanks to my Supervisor **Dr. Anil Sharma,** School of Computer Application, Lovely Professional University, Phagwara, Punjab for his help and guidance throughout my dissertation work. I have received an enormous amount of valuable advice and knowledge from him that helps me a lot in my research work. Thank you, **Dr. Anil Sharma**, for providing me a motivational support to carry my research work. I really feel short words to express my heartiest gratitude and sincere thanks to my parents, my professors for their support and all they have comprised for me during the tenure of my research work.

Signature of Candidate
Balrajpreet Kaur

Table of Contents

Chapter1: Introduction **Page no**

List of Tables

LIST OF FIGURES

CHAPTER-1

1.1 INTRODUCTION

Data mining is the process of finding patterns from large amount of data by applying some techniques. This is used as an analyzer for knowledge discovery in databases to be used in decision making process. Big organizations use it primarily for finding new ways to increase their profits and to minimize cost. Data mining analyze the data and helps to bring up the hidden factors so that useful patterns and information can be generated. As an instance, business organizations can analyze the customer's behavior toward specific product by analyzing the historical data and this will help the organization to find the changing behavior of the customer with the passage of time, like, to find the trends in change, to find the volume of change etc. These kinds of findings will definitely help any organization to take future decisions in relation to that product [1][2]. Data mining tools are the software which provide automatic implementation of data mining techniques on the data and provides user interface to apply machine learning algorithms [2]. These tools can handle huge amount of data and provide relevant results efficiently. Various tools are discovered with different parameters according to the need of users. The handling of data, user interface, missing values, finding error rate and many more parameters make these tools different from each other. These parameters can be increased or decreased according to the need of user. These tools are having features of handling complex as well as unstructured data [3]. Companies bought data mining tool to build their own customize mining solutions. Many Data Mining tools are available with their strengths and limitations in context to parameters like interfaces, algorithms, accuracy of results, mining techniques, data set size etc. These tools are further categorized into three categories i.e. Traditional data mining tools, Dashboards and Text Mining tools.

Traditional data mining tools mostly used by companies for business analytics purpose. These tools work on databases available with the company. There tools apply pre-defined algorithms on data for finding the invisible pattern and results. These tools provide broad data categories to generate readable reports. As an instance, a database of sales can display monthly sales results and reports with the help of traditional data mining tools. These tools are available both in Windows and Unix versions of operating systems and are mainly used for Online Analytical Processing (OLAP)[4]. Some of these tools are WEKA , R studio, Rapid Miner, SQL and D2K [5]. Dashboards are installed on computer to monitor database information and reflects the updates and changes onscreen regarding business information and performance. These are mostly used by companies which want to check its sales from historical point of view with the help of historical data i.e. Data Warehouse. Dashboards are easy to understand and it provide results in the form of charts and bar-graphs to provide overview regarding company's performance. All details related to profits and loss of company are visible to the manager on a single screen interface and the whole task is performed by dashboard features automatically. The leading dashboards provide the snapshot of actual performance of tools and also show the recent happenings [6]. The business intelligence dashboards are also known as enterprise dashboards [7]. These have the ability to pull the real time data from multiple sources. Oracle[6] and Microsoft[8] are among the leading vendors of business intelligence dashboards[10]. Text mining is analyzing the text to extract information that may be useful for particular purpose. It deals with natural language text and lexical usage to find useful information. Text mining tools easily access databases, scanned contents and include handling of structured and unstructured data. Text analytic software change unstructured data into numerical values so that it can link with structured data and find the result with traditional data mining tools. Apache mahout[9] is a tool which can handle structured and unstructured data. There are some text mining tools which are open sourced like orange[11] , NLTK[12] , Voyant[13] and ALchemy API[14]. IBM company build smarter Apps with ALchemy language[15] for semantic text mining[16] using Natural Language Processing[17]. This application help company to understand worlds conversation, reports and photos. These tools are progressively adding new features to satisfy the fast changing requirements of the user and to handle the data complexity in a better way. It is quite difficult to add all the features in one tool so there are different categories of tools introduced [2][18].

1.2 METHODS IN DATA MINING

1.2.1 Classification:- In this technique, the data is classified into different classes. One can represent classification in the form of maps. For example, countries can be classified into classes like developing countries and developed countries. There are so many classification algorithm used for classifying the data such as Decision trees, Naive Bayes, Generalized linear Model and support vector machine. The decision tree help in generating rules and make a tree . The Naive Bayes is mostly used for calculating probability . Support vector machine is used for both classification and regression. The classification having two testing techniques i.e, accuracy and confusion matrix The accuracy gives the correct predictions and confusion matrix shows the number of correct as well as incorrect predictions by the algorithm as compared with the actual classification in the test data. The classification provide application in the field of medical disease prediction as well as in biological data analysis for finding properties of various substances. Classification also used in multimedia data analysis such as photos, videos etc. The classification can perform functions on complex data set and it also provide application in social network analysis and customer interest in buying product.

1.2.2 Regression:- Regression method is used to map the relationship between two variables. Regression is also represented in the map form and we check the result by comparing the distance of data points from regression line. Profit, square footage , temperature ,sales and distance are predicted through regression. There are two types of regression linear and stepwise regression. Generalized linear model and support vector machine are the regression algorithm used for linear and non-linear regression. There are two formula's used for regression statistics i.e. Root mean square error(RMSE) and Mean absolute error.

1.2.3 Clustering:- In clustering, one perform the categorization of data into different categories. It provides the homogeneous data from huge amount of data. For example, one can make the cluster of metals which are affected by hydrogen and other cluster which didn’t have any effect. There are different methods used in clustering such as partitioning method, hierarchical method, density based method , grid based method , model based method and constraint based method. Partitioning method is the basic method used for clustering to group data. There are various applications of clustering in the field of marketing, biology ,fraud detection, similar land identification and many others.

1.2.4 Summarization:- Summarization is a method in which one make a compact description of any data. Summarization is done in the form of table. The summarization provides the relationship between different type of data sets. Summarization included both primitive and derived data. It provide relevant data from huge amount of data. There are two approaches to automatic summarization extraction and abstraction. Extractive method work on existing words, phrases or sentences in the original text to form the summary. Abstractive method use natural language generation techniques.

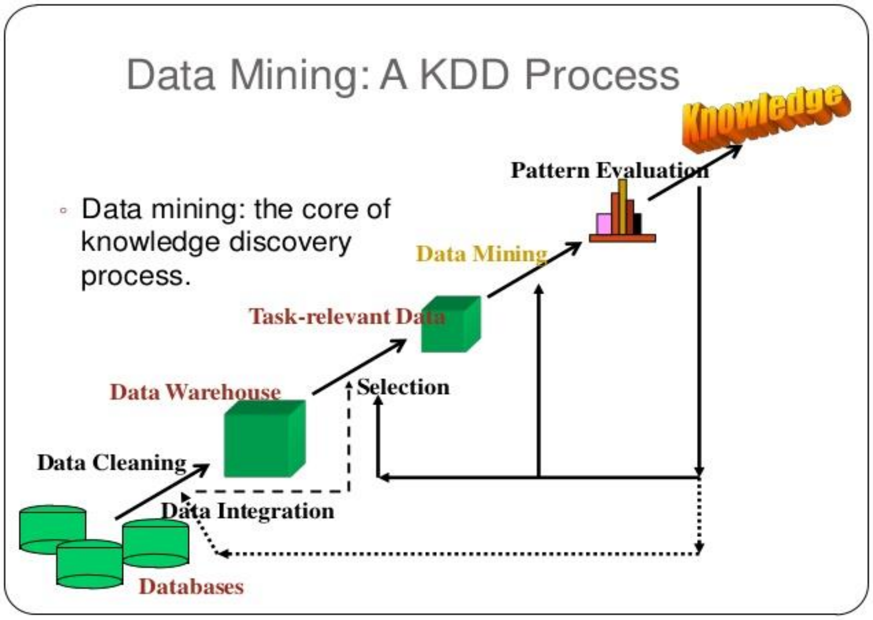
1.2.5 Introduction of some classifier

1. **Decision Tree :-** Decision tree is a procedure for classifying the data on the basis of attributes. Decision tree can analyze large amount of data so it is a data mining application. There is no requirement of domain knowledge needed in this algorithm. The output is represented in the form of tree which helps in understanding the classification easily. It breaks down the data into smaller and smaller sets and the associated tree is incrementally developed. The final tree contain decision node and leaf node. Decision tree can handle numerical as well as categorical data. There are some algorithm for building decision tree such as ID3. ID3 uses entropy and information gain to make decision tree.
2. **K-Nearest Neighbor :-** K-NN is a non-parametric method for classification as well as regression. In K-NN classification an object is classified by majority of its neighbor. The object is assigned to the most common class among its K nearest

neighbor. There is a process of parametric selection in this algorithm. The larger the value of K reduce the effect of noise in the result of classification. The K value is selected by heuristic techniques. The accuracy in K-NN can be improved by two algorithms i.e. large margin nearest neighbor and Neighborhood component analysis. This algorithm provide accuracy as well as error rate.

3. **Naive Bayes:-** Naive Bayes is a classification technique based on the Bayes Theorem. This classifier assumes that the presence of a particular feature that is unrelated to the other feature in the class. Naive Bayes is useful for large dataset and give higher accuracy for sophisticated classification method. There are some application of Naive Bayes algorithm such as real time prediction, multiclass prediction and text classification. Naive Bayes uses the probability concept.
4. **Support Vector Machine:-** Support vector machine is a supervised learning model to analyze data for classification and regression . SVM construct a hyper lane in a infinite dimensional space. The support vector and data points that is within the decision surface. There are some application of this model such as text categorization , classification of images and image segmentation.

1.3 ARCHITECTURE OF DATA MINING


Data Mining: A KDD Process

◦ Data mining: the core of knowledge discovery process.

Figure 1.1: Architecture of data mining

I. In first step we extract data, then transform and load the data into data warehouse system.

II. We store the data and manage it in multidimensional databases.

III. After that we analyze the data and make the patterns from the data by applying some techniques.

IV. We get the information and knowledge from the data.

1.4 APPLICATIONS OF DATA MINING

1.4.1 Sales/ Marketing: Data Mining provides help in analyzing the past sale data and during this analysis this can be seen that how sales patterns are changing with the passage of time which can be further very helpful for an organization plan their sales strategies and marketing if the product as well. For example, the stock data is divided into three cluster i.e. Dead stock (DS), slow moving(SM), fast moving(FM). The dead stock contains record of those product who are less sold by customer . Slow moving having medium sales stock cluster and fast moving having large selling quantity. There is a new technique introduce i.e. most frequent pattern for finding clusters.

1.4.2 Banking and Finance: Mining has a great role to play in the area of banking and finance. A new type of crime is extending i.e. Cyber Crime which is hitting very badly the banking and finance sector. Data mining is helping to develop applications which can be used to detect credit card frauds. Transaction frequency can also be analyzed through mining techniques through which the loyalty of customer can also be predicted. Fraud detection through data mining techniques such as classification, clustering , prediction and many more. There are 26 data mining techniques applied to the detection of financial fraud. Data mining technique used in type of fraud such as bank fraud , Insurance fraud, automobile insurance fraud and many more.

1.4.3 Biological Field: Data mining provide major application in medical field to find the correlation among different drugs. It is also helpful to diagnose disease a patient is suffering from by analyzing symptoms correlation. For example data mining help doctors to find precision medicine. The association technique are used for finding relation between genetic markers and disease risk. This help in genetic analysis of human disease. There is one more example of salt intake relation with increase in blood pressure. There are different effect of salt intake on humans.

1.4.4 Other applications: - Data mining can also be used in geosciences, simulations and other engineering fields. It is the most important tool for predictions.

1.5 VARIOUS TOOLS IN DATA MINING

1.5.1 **Rapid Miner**: Rapid Miner is an open source data mining tool. The name of Rapid Miner now change into Rapid Miner Studio. The Rapid Miner provide application for data analysis, data transformation , visualization, predictive analytics and business analytics. Rapid Miner is a package of machine learning algorithms. The Rapid Miner also used in sentiment analysis, churn reduction and predictive maintenance. For Rapid miner user don't need any programming language. The user can easily select charts and visualization. It is having advanced analytics by giving interface for big data handling.

1.5.2 WEKA: Weka is a collection of machine learning algorithm. The main advantage of WEKA is that it is directly applied to dataset. Weka is written in Java. Weka is having interfaces such as Explorer, Experimenter, knowledge flow and simple CLI. The user mostly use explorer part to experiment their dataset and finding the result. Experimenter interface is used for testing and evaluating. Knowledge discovery interface provide KDD process and Simple CLI is a command interface without GUI. Weka provides pre-processing tools, classification, clustering, association and visualization. There are different algorithm provided in each categories . In classifier output the results are in the form of numerical values.

1.5.3 Orange: Orange is an open source for data visualization and data analytics for expert and novice. It provide colorful graphical interface and workflows. It is python scripted. The orange widgets made it different from other tools. The user can make workflow between different widgets according to their needs. The scatter plot of orange tool is very clear and colorful. The orange is having machine learning algorithm and it is mostly used for bioinformatics and text mining . It contain machine learning add-ons for bioinformatics and text mining. Its having scoring feature and cross validation process. This tool provide various visualization methods . The various applications of orange are in field of Aerospace, Agriculture, Chemical and business analytics.

1.5.4 KNIME: The KNIME GUI provides a platform to create data flows, data analysis and review the result and models. KNIME is written in JAVA and built on Eclipse. There are different applications of KNIME such as image mining, text mining , time series analysis and many more. It is used for data processing which includes extraction, transformation and loading. It include machine learning and used for business intelligence and financial data analysis. KNIME is mostly used for text mining process KNIME also provide clustering execution and the interface provide user a platform to make own data flow and applied techniques on dataset.

1.5.5 NLTK: It is natural language tool kit. It basically used for sentiment analysis and various languages processing task. It written in python. It can divide the sentences into token and also provide the tokenization. It is mostly used for analyzing the customer behavior.

1.6 CHALLENGES IN DATA MINING

1.6.1 Poor quality and noisy data: The poor quality of data having missing values and various complication in data. Sometimes it's very difficult for a data mining tool to handle these type of data and provide efficient results.

1.6.2 Lack of understanding: The data which is difficult to classified and having no sense is a complex data. Sometimes it's really difficult to handle messy data without understanding it.

1.6.3 Data variety: The heterogeneous data is containing data from various source and its having different varieties of data. the data mining tool facing problem in handling these type of data.

1.6.4 Huge amount of data set : The biggest challenge is the handling of huge amount of data such as big data . The huge amount of data is stored from various sources which make data more complex.

1.6.5 Complex questions related to data: If a user doesn't get the answer he/ she want to find from the data then it become a challenge. Sometimes the data mining doesn't provide the result according to the user needs.

CHAPTER-2

2.1 LITERATURE REVIEW

Data mining is the process of exploring unexplored patterns from huge databases. It provides a great support to business world and academia . Data mining tools provide interface to get data and to produce some interesting patterns out of it which are further useful to attain new knowledge. In [19] the author gives introduction about the data mining and data mining techniques. The data mining is a part of knowledge discovery. The data mining techniques such as association rules, classification, regression and cluster analysis are described. The author provide the mathematical framework of each technique. The paper is having the mathematical formula and detail of each technique with equations. The author discuss about the complexity of each algorithm with datasets. This paper help in understanding algorithm and make changes in improving the efficiency of algorithms. In [20] the author give introduction about data mining and knowledge management. The authors gives description about the data mining techniques with examples. There is a framework in the paper which show the knowledge management integrated with knowledge management cycle. This framework include knowledge capturing , knowledge acquisition. The authors also give introduction about knowledge management tools and technologies. There are different type of knowledge types described in the paper such as healthcare system domain, construction industry domain, financial domain, business domain. The paper describe about the applications of data mining used in knowledge management such as classification, clustering and dependency model. Each data mining technique is described with various algorithms. In [21] the author give introduction about data mining and history of data mining. The paper shows the various applications of data mining in various field. The main aim of data mining is to discover knowledge from databases. The author describe about educational data mining, online learning, student learning behavior. The data mining provide application in medical field also such as finding correlation between different medicines and also diagnose disease from symptoms. Data mining also provide application in mechanical engineering by analyzing the properties of material and also help in finding complex answers.

In [22] there are various classification algorithm discussed such as decision tree algorithm, J48 algorithm and K-means. The author give pros and cons of WEKA and Orange . The comparison of data mining algorithm provide results i.e. KNN give less accuracy as compared to decision tree and Bayesian. After reading the data mining and data mining techniques paper we focus on data mining tools performance and we read many papers regarding the comparison of data mining tools.

In [23], the author gave information about WEKA. WEKA is a open source it is implemented in java language. WEKA is used for implementing the various data mining methods. The most positive point in WEKA is its maintenance and modification feature. In [24], the author provide orange tool working for text mining. It provides sieve diagrams and parallel diagrams. It is implemented in C++. It is good in run time and also decrease the error rate. In [25], the author gives introduction about KNIME architecture and its functionalities. The KNIME tool provides an interface having analysis flow. KNIME is written in java . The KNIME framework is having three parts: visual framework, modularity and easy expandability. The KNIME also have repository which having mining algorithm ,data transformation and data input and output . The KNIME tool also integrates with different tools such as WEKA , R-project and Jfree chart. In [26], the authors give introduction about NLTK (Natural Language Tool Kit).This tool is scripted in python. NLTK is basically used for text mining. The basic working of NLTK is tagging, grammar and tokenization of word. Hidden Markov model and other language modeling is introduced in NLTK. In [27], the authors explains the working of AR Miner tool which is based on Association rule. This tool helps in decision making process. The functions include data preparation and mining association rules included negative items. The author also gives example of application of AR Miner. Author further emphasized that patterns can be generated from the negative items also. In [28], the author give introduction about rough set theory. The rough set theory is used for extracting decision rules from data sets. The author gives the example of rough set theory. In the example author taken four features and according to the if-else rules various results are drawn. In [29], the author introduces the evolutionary methods and tools for classification of data. KEEL is made in java language. It provide good interface to the user. It consist of new analysis model and better than other tools. In [30], the author explains three concepts in the paper. First, it shows the comparative study of data mining tools. Second, it explains about different challenges in data mining tool. Third, it explains the advantage of agent with data mining tools. The comparative study is based on portioning of dataset, scaling, selection, parameter optimization of machine type learning .the agent provide intelligence in the system. Different type of agents are coordinate agent, clean agent, Reduction agent and transformation agent. These agents help in overcome the challenges.

In [31], the author explains Tanagra, is a tool used for different operations and it is used for diagram making. Tanagra is better in overcome the error rate. Tanagra is good classifier and can handle huge amount of data. In [32], the author give information regarding Data mining tools for doing data mining process and finding new patterns. The SPSS tool is used for finding regression and correlation. The author categorize he tool into nine types based on the suites, business intelligence packages, mathematical packages, integration packages and other libraries and solutions. In [33], the author take 3 tools i.e. WEKA, Tanagra and Clementine. The author test four healthcare Dataset on these tools. In the result the author took two parameters : accuracy rate and error rate. Different techniques are applied on dataset to get the accurate rate and error rate percentage. In other paper [34], the comparison of data mining tools is based on nine different types of data. The author used six algorithms for classification such as naïve Bayes, support vector machine(SVM), zero rule, one rule and decision tree. The test conducted results that all the tools are good . The performance of the tool is decrease and increase on the basis of datasets we are using. The WEKA tool is better in classification as compared to other tools. The other comparison is done in[35], the author used two data mining tools i.e. Tanagra and WEKA. The dataset consist of 100 patients from research centre. The aim of data mining is to find out the relationship between diabetic patient and kidney failure. Both the tools are used as classifier and C4.5 algorithm is used for making decision tree. From the result it is found that the error rate of Tanagra is 11% and WEKA error rate is 25%. From above papers the comparison result show that WEKA tool is better than other tools in the field of classification technique. The time taken by WEKA is less as compared to other tools but error rate is little high than Tanagra.

In[36],the author check the accuracy about complaint detection task. The authors take three data set in which they had so many complaint review and non-complaint reviews. The author chooses five algorithm for test, one rule, conjunctive rule ,ridor, RIPPER and PART. The performance measurements are taken on the basis of accuracy values. In the result, the 75% accuracy is there as compared to other algorithms. The process also contain the removal of unigrams. In [37], the author introduces four parameters for calculating the performance of data mining tools. Performance factors include computational performance, functionality, usability and ancillary task support. These all factors performance depends upon the quality of data. Computational performance indicates the tool ability to handle data in varieties. Functionality describes how to solve different types of data mining problems. Usability indicate how tool is used by user in efficient way with functions. Ancillary task support provides functions like data cleaning, transformation, visualization and other task. The author gives the result table in which they show the ANOVA test. In [38] the comparison is based on another two criteria between the five data mining tools The result analysis is based on two criteria i.e. first the user evaluation analysis and second is the technical evaluation analysis. Different types of parameters are considered for getting results. There are some participants invited for fulfill the evaluation criteria. The participants give the rating from 0 to 10 to find the comparative result between these tools. In results the author mentioned that the R studio is fail to impress the participants while the other tool give better results. KNIME and Rapid Miner get better score than other tools. Weka tool is lacking the interface and feel outdated. Each tool have their advantages and disadvantages the score criteria shows how much people feel comfortable with these tools. These two papers show the comparison based on user review as well as the working of tool. In [39] , the author take three different data mining classification method for comparison . The author used breast cancer dataset for finding the result. The Weka tool is used for finding the result of classification . The three algorithm used as classifier are Decision tree, Bayes classification and K- nearest neighbors . In this paper the description about these algorithm are defined. The result table is shown in the paper. There are different parameters taken for finding the comparative analysis. The Bayes Classification is best as it takes the less time i.e. 0.02 sec and also give accuracy of 95.9943 in classification. On the basis of parameters Bayes Classification is best in comparison with other classifiers. Same as the classification algorithm comparison there is a clustering algorithm comparison in [40], the author give introduction about different clustering algorithm present in WEKA tool. The authors make a comparison table based on four things i.e. time to build a model, cluster instances, squared errors and log likelihood by using WEKA tool. They used the data of emit software repositories. In the result the author show that k-means algorithm is the best algorithm for clustering because it take less time to build a model and gives efficient result as compared to other algorithm.

The above two papers[41][42] the result is based on numerical value but in [41], the author focuses on four data mining algorithms K-NN, Naive Bayes classifier , Decision tree and C4.5. The author did the comparative analysis on the basis of theory, advantages, disadvantages and applications. The decision tree is based on if-then rules. The KNN is the oldest algorithm . The naive Bayes is Simple and easy to understand. C4.5 algorithm is mostly used for real life problems. C4.5 algorithm provide decision tree for visualizing the classification. In [42], the author gives the comparison result of three data mining tools and a new framework DMPML(Data Mining Preparation Markup Language). The DMPML can stores directives and codified data in an XML document. The result is based on two parameters the creating of directed graphs and time processing in output of data. The DMPML spend less time in creating directed graph as compared to three data mining tools but it take more time in processing the output data as XML document. The DMPML requires less user interaction as compared to other data mining tools. The results of this paper gives appropriate difference between these tools and DMPML framework. In [43] the author compare three tools WEKA, Rapid Miner and KNIME on the basis of parameters i.e. developer, programming language, released date, license, availability, current version, areas , usability, compatibility with database, platform supporting, flexibility, visualization and GUI. In [44] the author used different dataset and check the performance of K-means data clustering and Naive Bayes data classification method. The author use attribute selection technique for the improvement in accuracy by 3.49% and 2%. The parameters taken for checking the results are time, accuracy , precision and recall. The author shows the graph which give the accuracy and many others results value with improvements. For further application the analysis is on the basis of medical dataset and communication dataset[45][46]. In [45], the author take patient dataset for checking the best classification method for medical decisions. The author use WEKA tool for classification results. There are 10 classifier used for comparing the results . The author had taken 8 parameters for comparing the classification results. The parameters are TPrate, FPrate, precision, recall , F-measure, ROC area and time taken . After comparison Bayes Net give the best classification results with TPrate and other parameters. In [46] the author take 2 tools KNIME and Rapid

Miner and check their accuracy based on some experimental model. The author provide KNIME tool workflows and results in the form of Pareto chart and bar charts. The experiment design is to evaluate the data mining tool based on both quantitative and qualitative approaches . The author also analyzed tool on the basis of workload size.

2.2 RESEARCH GAP

According to the literature survey, the comparison of data mining tools is done on various parameters such as developer, programming language, released date, availability, usability, compatibility, platform supporting, visualization and versions[18][19][21][23][24][20]. Most of the authors done comparison on the above parameter or they do have classification on some specific dataset [10][12][19][24][28]. But, there is very limited literature available which has results in the form of numeric's. There is hardly any paper to compare data mining tools on the basis of factors. This gap motivated us to explore and find new possibilities in this direction.

2.3 PROBLEM FORMULATION

"To Find The Factors That Affect The Performance Of Data Mining Tools"

CHAPTER-3

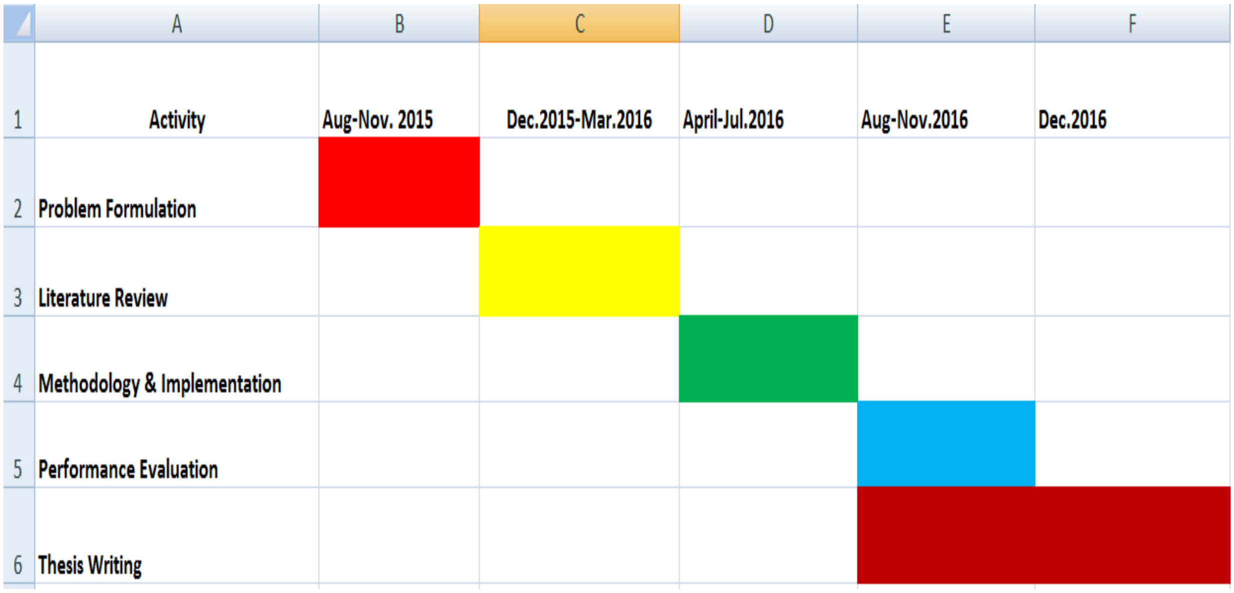
3.1 SCOPE OF STUDY

Data Mining tools are the most reliable tool for data analysis. Now a days, most of the organizations used data mining tools for analyzing their data and for getting information from huge amount of data. The data mining tools provides automatic functionality for data analysis. From literature survey we found there are many issues in data mining tools such as scalability, performance issue, handling complex data and many more . From these issues we consider the performance issues. From literature survey we found the various data mining tools which are easily available i.e. KNIME, WEKA, Orange, Rapid Miner, CMSR, SPSS, Matlab. From these tools we took three tools - Orange, WEKA and Matlab. For comparing these tools we consider same algorithms present in these tools and compare them with existing parameters i.e. time taken, correctly classified percentage and incorrectly classified percentage. For testing two datasets are used . From comparing result we found there are variation in results and type of dataset also effect the performance of data mining tool.

3.2 OBJECTIVES

1. Testing of different input data sets on tools .
2. Comparative analysis of data mining tools.
3. Comparison of factors affecting the performance of different data mining tools.

3.3 TIMELINE



	A	B	C	D	E	F
1	Activity	Aug-Nov. 2015	Dec.2015-Mar.2016	April-Jul.2016	Aug-Nov.2016	Dec.2016
2	Problem Formulation	🟥				
3	Literature Review		🟨			
4	Methodology & Implementation			🟩		
5	Performance Evaluation				🟦	
6	Thesis Writing				🟥	🟥

3.4 RESEARCH METHODOLOGY

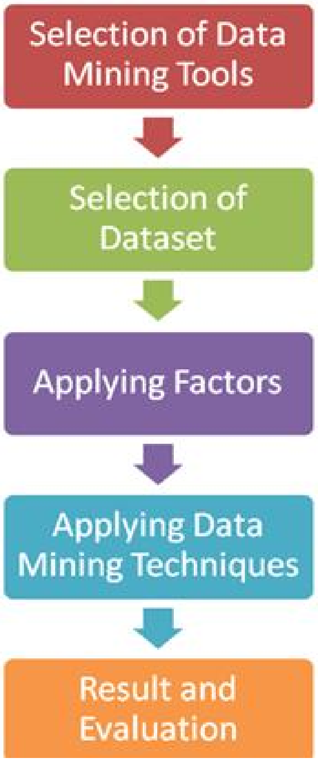

Selection of Data Mining Tools
Selection of Dataset
Applying Factors
Applying Data Mining Techniques
Result and Evaluation

Figure 3.1: Research methodology

We have take some data mining tools and test different data set. The different type of datasets are considered for testing the tool performance . The data set that is used for testing can be data with missing value, structured data set, multivariate dataset, unstructured data and many more. The different data mining techniques such as classification with different algorithm is used for comparison. We can compare the tool performance on the basis of algorithm used for each data mining technique with same dataset. After testing the dataset we can validate the result by checking accuracy, time ,handling of data and other factors . These factors is the output.

CHAPTER-4

4.1 EXPERIMENTAL WORK:-

In experimental work we selected the data mining tools and get them installed. We took data mining tools - Orange, WEKA, Matlab. After that we applied data classification technique on datasets to find the output of each data mining tool. In our research we tested different classification algorithms on two datasets and checked the behavior of data mining tools in accordance to them. The output of data mining tool highlighted the comparative analysis of common parameters in different tools.

4.2 RESULTS AND DISCUSSIONS

Dataset:- We have taken two dataset . The first dataset is adult dataset . This dataset contain the census data of 1994. The prediction task to determine the person over 50K a year. This dataset contain the missing value.

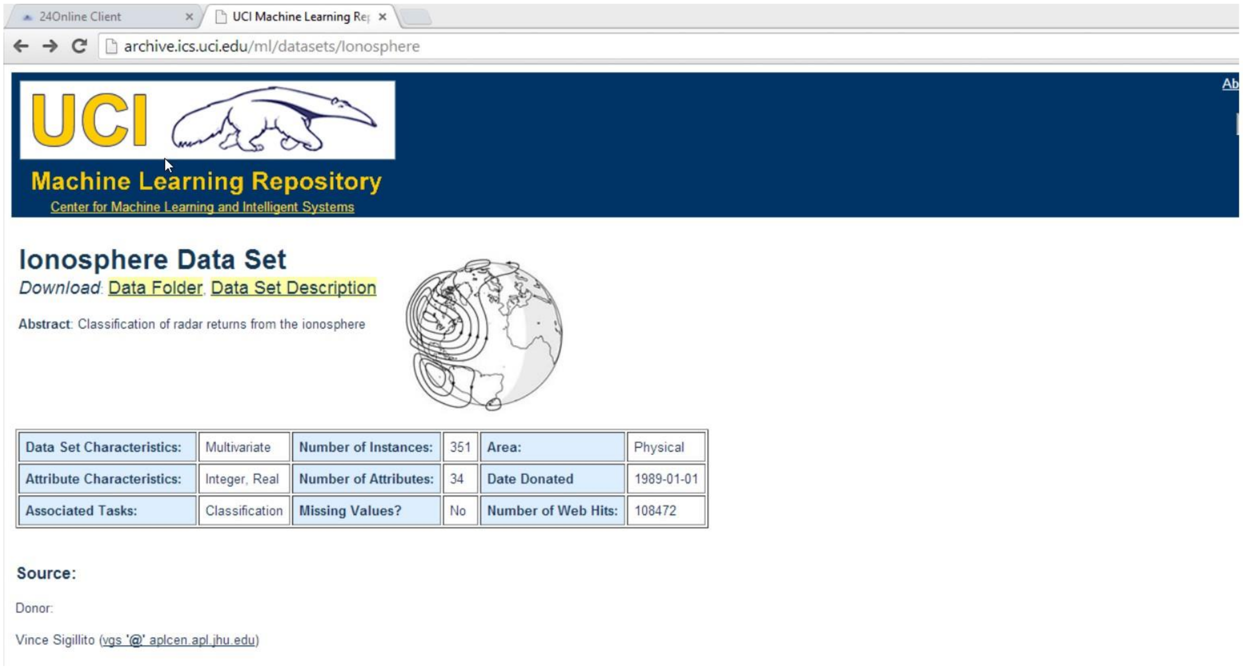
The attributes in the dataset are:-

a. age: continuous.
b. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
c. fnlwgt: continuous.
d. education: Bachelor's, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
e. education-num: continuous.
f. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
g. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

h. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
i. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
j. sex: Female, Male.
k. capital-gain: continuous.
l. capital-loss: continuous.
m. hours-per-week: continuous.
n. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

THE SCREENSHOT OF THE DATASET

THE SCREENSHOT OF SECOND DATASET


24Online Client
UCI Machine Learning Re
archive.ics.uci.edu/ml/datasets/Ionosphere
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Ionosphere Data Set

Download: Data Folder, Data Set Description

Abstract: Classification of radar returns from the ionosphere

Data Set Characteristics:	Multivariate	**Number of Instances:**	351	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	34	**Date Donated**	1989-01-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	108472

Source:

Donor:

Vince Sigillito (vgs '@' aplcen.apl.jhu.edu)

The second dataset is the ionosphere dataset. The dataset contain 34 attributes. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not their signals pass through the ionosphere.

a. Number of Instances: 351
b. Number of Attributes: 34 plus the class attribute. All 34 predictor attributes are continuous
c. Attribute Information: All 34 are continuous, as described above. The 35th attribute is either "good" or "bad" according to the definition summarized above. This is a binary classification task.
d. Missing Values: None

We have taken three tools i.e. Orange, WEKA and Matlab.

The results of WEKA with first dataset i.e. adult dataset. Naive Bayes, J48, Weight handler wrapper are the algorithm used .

SCREENSHOTS

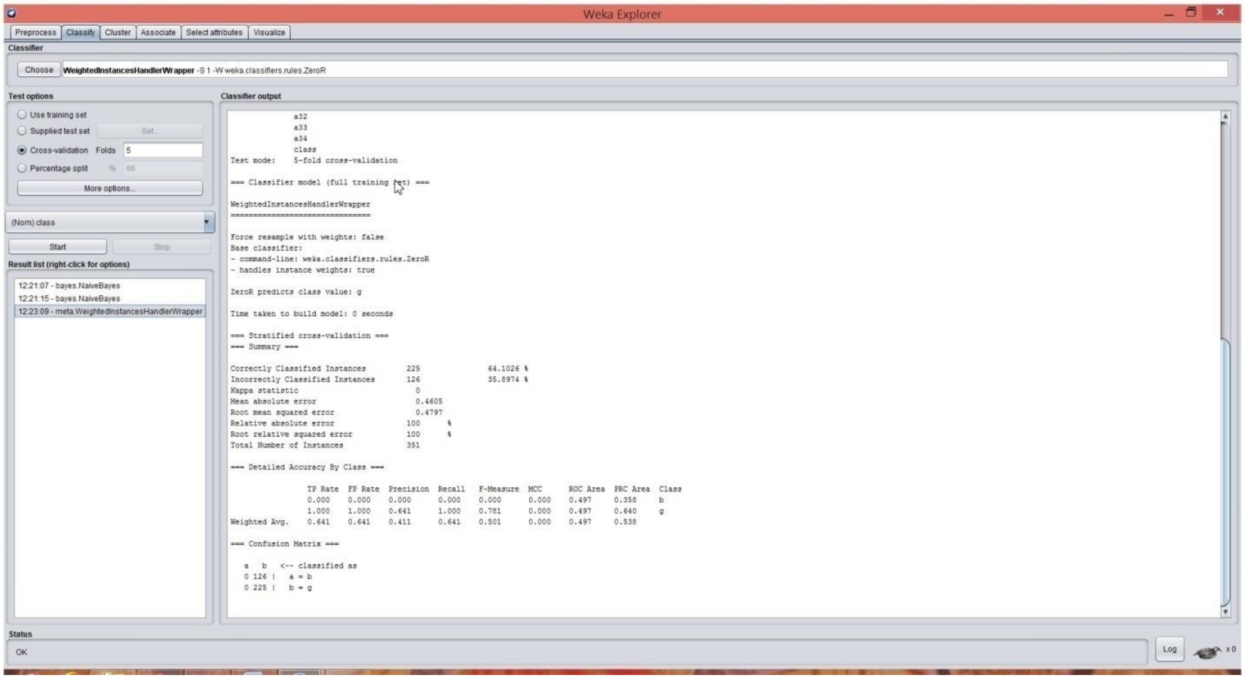
THE SCREENSHOT OF RESULTS IN WEKA WITH NAIVE BAYES ON ADULT DATASET.

THE SCREENSHOT OF RESULTS IN WEKA WITH J48 ON ADULT DATASET.

SCREENSHOT OF WEKA WITH IONOSPHERE DATASET.

SCREENSHOT OF RESULT IN WEKA WITH NAIVE BAYES ON IONOSPHERE DATASET

SCREENSHOT OF RESULT IN WEKA WITH KNN ON IONOSPHERE DATASET.



Choose: WeightedInstancesHandlerWrapper -S 1 -W weka.classifiers.rules.ZeroR

● Cross-validation Folds 5

% 66

(Nom) class

12:21:07 - bayes.NaiveBayes
12:21:15 - bayes.NaiveBayes
12:23:09 - meta.WeightedInstancesHandlerWrapper

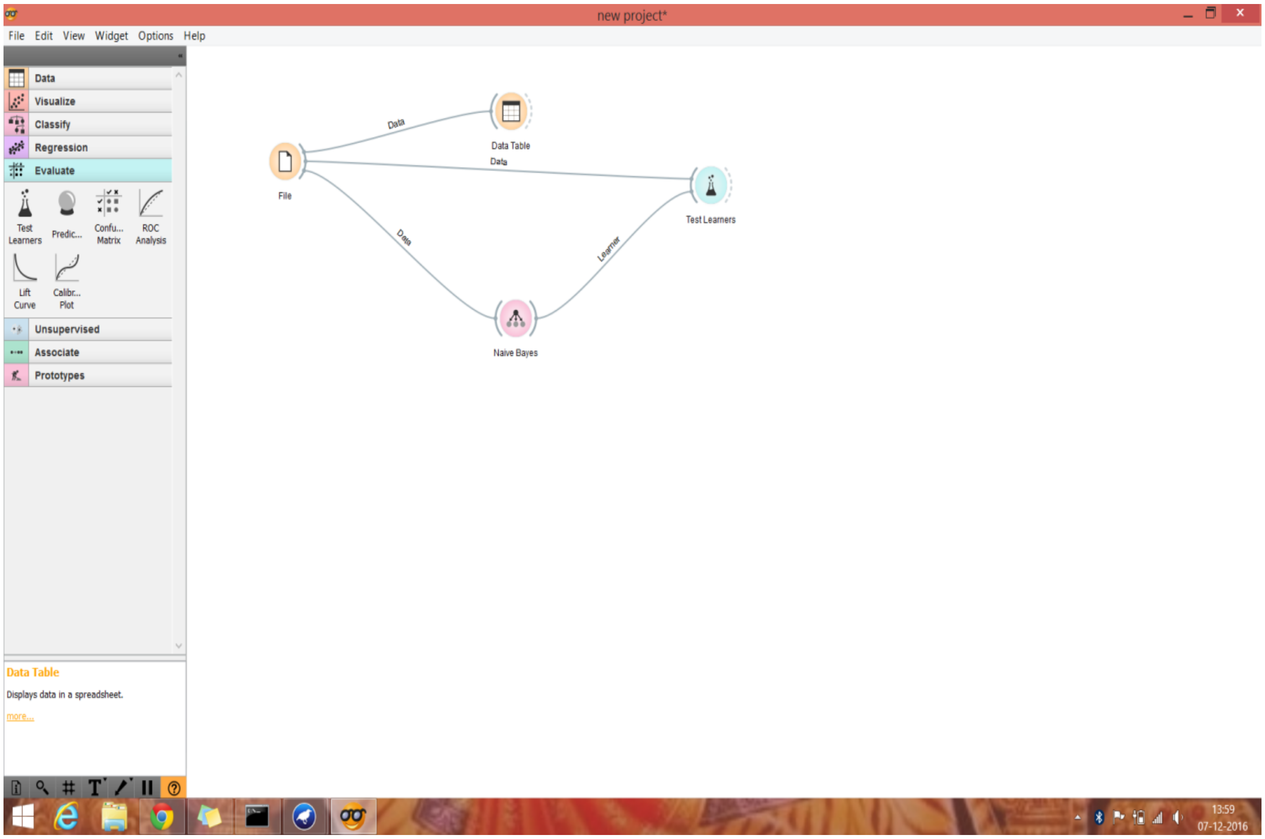
Status: OK

```
              a32
              a33
              a34
              class
Test mode:    5-fold cross-validation

=== Classifier model (full training set) ===

WeightedInstancesHandlerWrapper
===============================

Force resample with weights: false
Base classifier:
- command-line: weka.classifiers.rules.ZeroR
- handles instance weights: true

ZeroR predicts class value: g

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         225               64.1026 %
Incorrectly Classified Instances       126               35.8974 %
Kappa statistic                          0
Mean absolute error                      0.4605
Root mean squared error                  0.4797
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              351

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.497     0.358     b
                 1.000    1.000    0.641      1.000    0.781      0.000    0.497     0.640     g
Weighted Avg.    0.641    0.641    0.411      0.641    0.501      0.000    0.497     0.538

=== Confusion Matrix ===

   a   b   <-- classified as
   0 126 |   a = b
   0 225 |   b = g
```

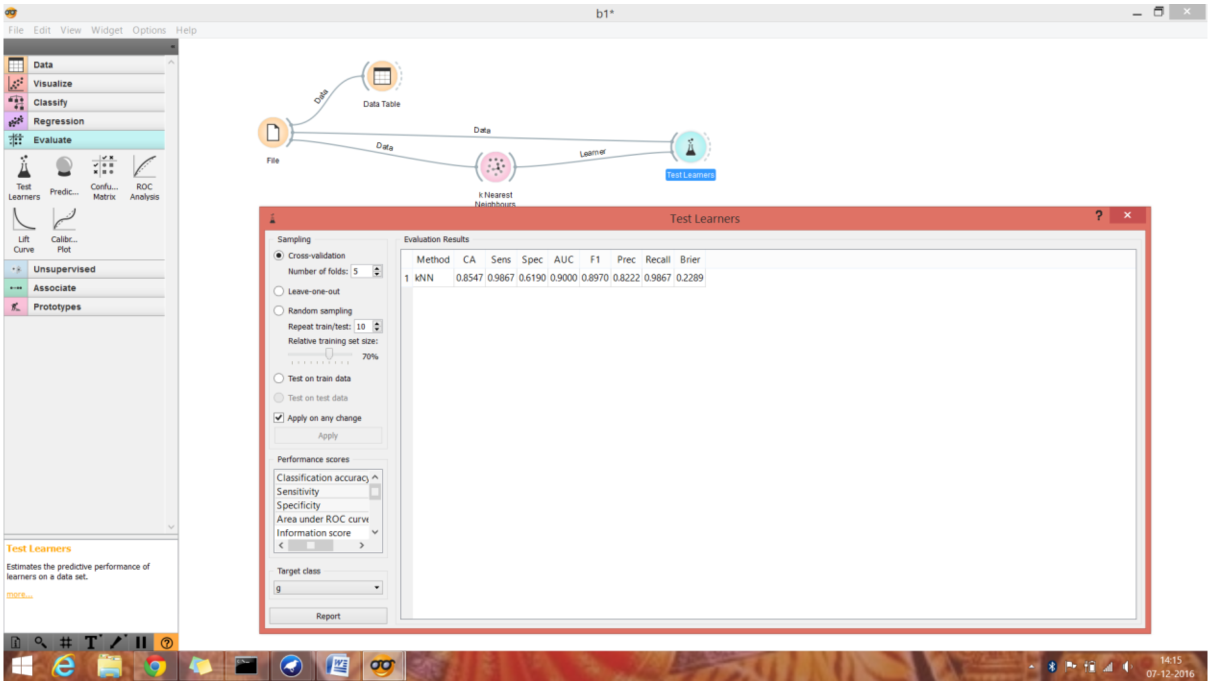
SCREENSHOT OF RESULT IN WEKA WITH SVM ON IONOSPHERE DATASET

SCREENSHOT OF RESULT IN WEKA WITH J48 ON IONOSPHERE DATASET.

ORANGE RESULTS WITH IONOSPHERE DATASET



SCREENSHOT RESULT WITH KNN ON IONOSPHERE DATASET



Test Learners

Estimates the predictive performance of learners on a data set.

more...

Sampling

● Cross-validation

Number of folds: 5

○ Leave-one-out

○ Random sampling

Repeat train/test: 10

Relative training set size: 70%

○ Test on train data

○ Test on test data

☑ Apply on any change

Apply

Performance scores

Classification accuracy
Sensitivity
Specificity
Area under ROC curve
Information score

Target class: g

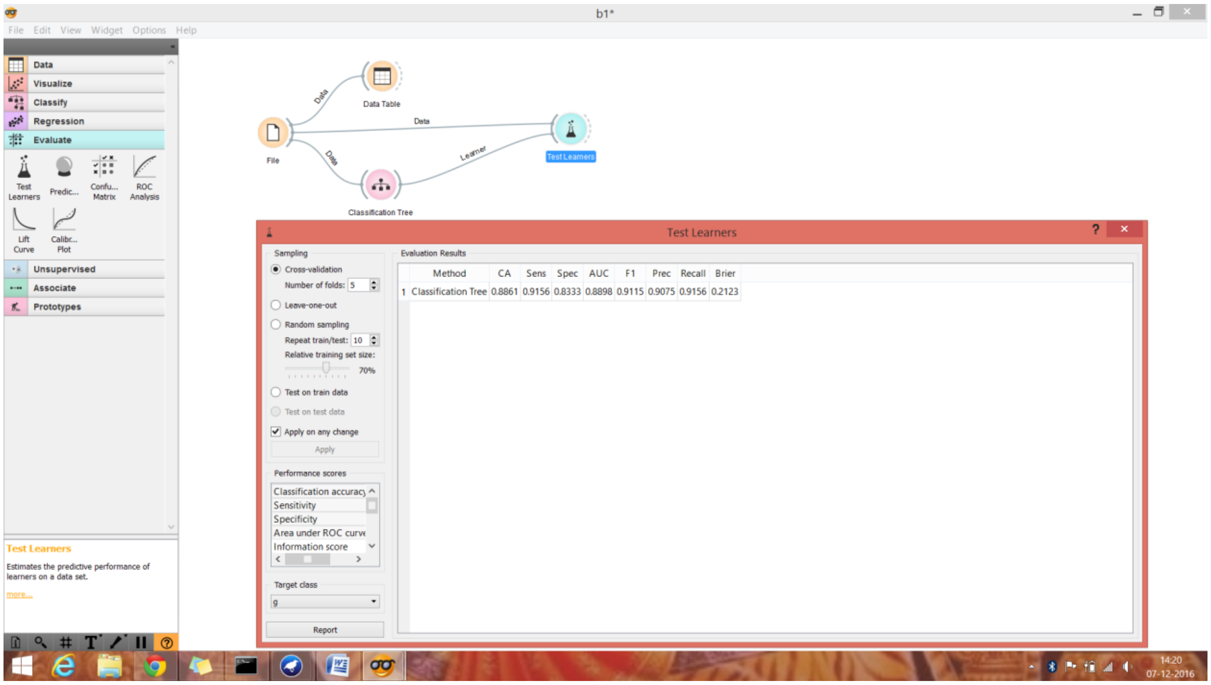
Report

Evaluation Results

	Method	CA	Sens	Spec	AUC	F1	Prec	Recall	Brier
1	kNN	0.8547	0.9867	0.6190	0.9000	0.8970	0.8222	0.9867	0.2289

SCREENSHOT RESULT WITH SVM ON IONOSPHERE DATASET

SCREENSHOT RESULT WITH CLASSIFICATION TREE ON IONOSPHERE DATASET



Test Learners

Estimates the predictive performance of learners on a data set.

more...

Test Learners

Sampling

- ● Cross-validation
  - Number of folds: 5
- ○ Leave-one-out
- ○ Random sampling
  - Repeat train/test: 10
  - Relative training set size: 70%
- ○ Test on train data
- ○ Test on test data
- ☑ Apply on any change

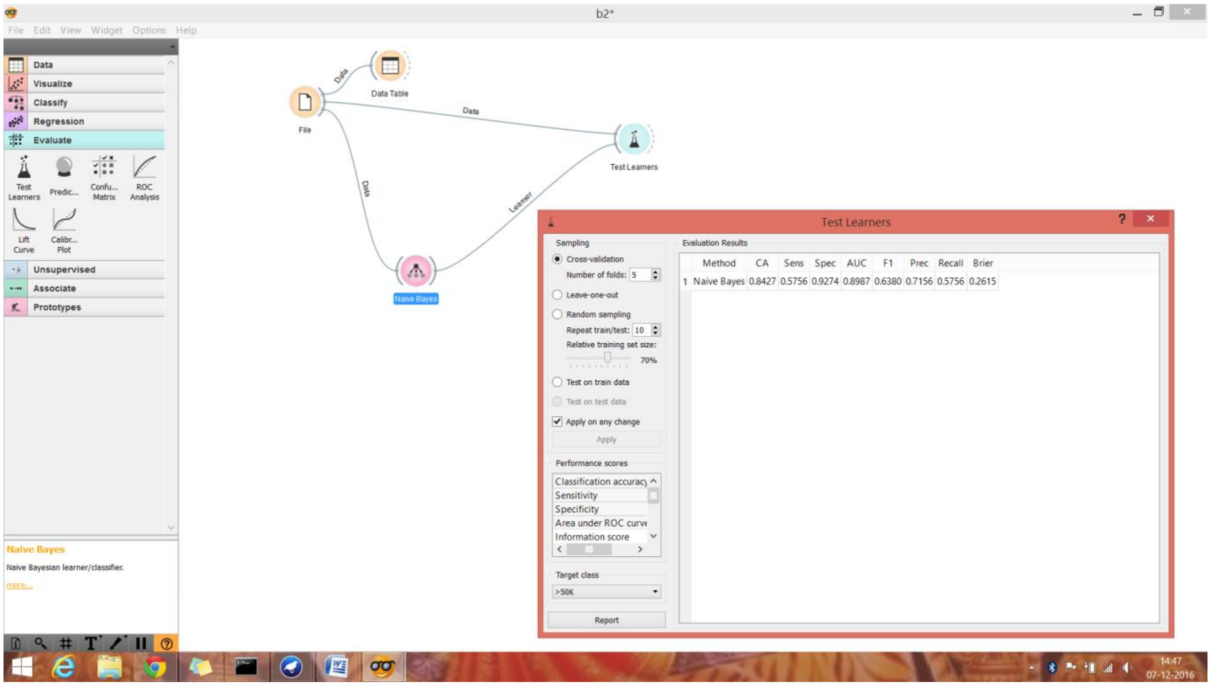
Performance scores: Classification accuracy, Sensitivity, Specificity, Area under ROC curve, Information score

Target class: g

Evaluation Results

	Method	CA	Sens	Spec	AUC	F1	Prec	Recall	Brier
1	Classification Tree	0.8861	0.9156	0.8333	0.8898	0.9115	0.9075	0.9156	0.2123

SCREENSHOT OF RESULT ON ADULT DATASET WITH NAIVE BAYES



Naive Bayes

Naive Bayesian learner/classifier.

more...

Test Learners

Sampling

- ● Cross-validation
  - Number of folds: 5
- ○ Leave-one-out
- ○ Random sampling
  - Repeat train/test: 10
  - Relative training set size: 70%
- ○ Test on train data
- ○ Test on test data
- ☑ Apply on any change

Performance scores: Classification accuracy, Sensitivity, Specificity, Area under ROC curve, Information score

Target class: >50K

Evaluation Results

	Method	CA	Sens	Spec	AUC	F1	Prec	Recall	Brier
1	Naive Bayes	0.8427	0.5756	0.9274	0.8987	0.6380	0.7156	0.5756	0.2615

SCREENSHOT OF RESULT ON ADULT DATASET WITH CLASSIFICATION TREE

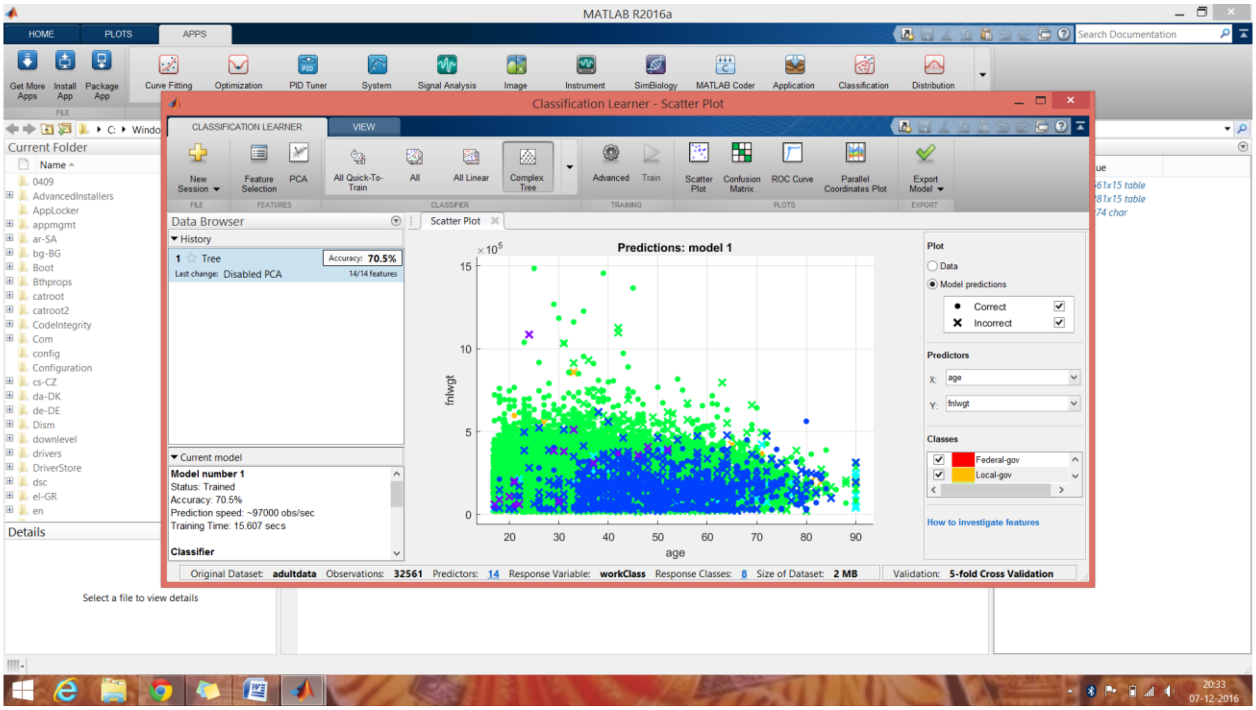
SCREENSHOT OF RESULT ON ADULT DATASET WITH KNN

MATLAB RESULT ON IONOSPHERE DATASET WITH J48

MATLAB RESULT ON IONOSPHERE DATASET WITH NAIVE BAYES

MATLAB RESULT ON IONOSPHERE DATASET WITH KNN

MATLAB RESULT ON IONOSPHERE DATASET WITH MATLAB RESULT ON ADULT DATASET WITH J48



Data Browser

History

1 Tree Accuracy: 70.5%
Last change: Disabled PCA 14/14 features

Current model

Model number 1
Status: Trained
Accuracy: 70.5%
Prediction speed: ~97000 obs/sec
Training Time: 15.607 secs

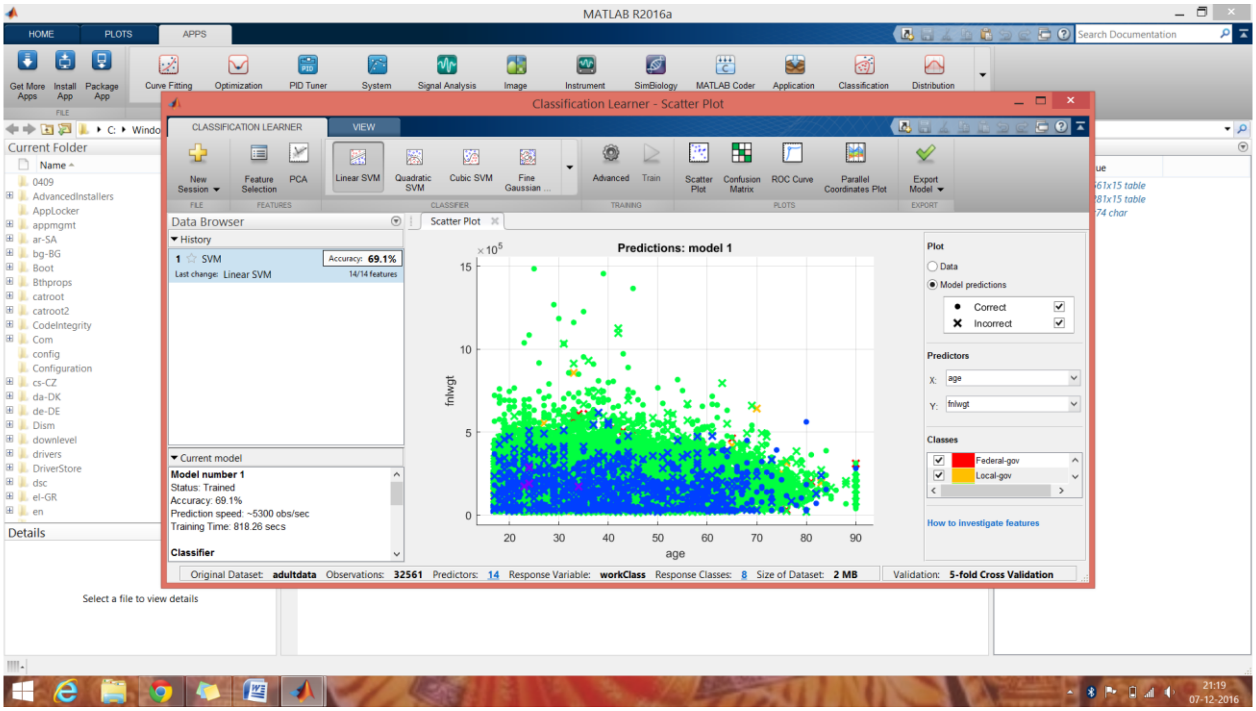
Classifier

Plot: ○ Data ● Model predictions — ● Correct ☑ ✕ Incorrect ☑

Predictors: X: age Y: fnlwgt

Original Dataset: **adultdata** Observations: **32561** Predictors: **14** Response Variable: **workClass** Response Classes: **8** Size of Dataset: **2 MB** Validation: **5-fold Cross Validation**

MATLAB RESULT ON ADULT DATASET WITH SVM



Data Browser

History

1 SVM Accuracy: 69.1%
Last change: Linear SVM 14/14 features

Current model

Model number 1
Status: Trained
Accuracy: 69.1%
Prediction speed: ~5300 obs/sec
Training Time: 818.26 secs

Classifier

Plot: ○ Data ● Model predictions — ● Correct ☑ ✕ Incorrect ☑

Predictors: X: age Y: fnlwgt

Original Dataset: **adultdata** Observations: **32561** Predictors: **14** Response Variable: **workClass** Response Classes: **8** Size of Dataset: **2 MB** Validation: **5-fold Cross Validation**

4.3 RESULT TABLE

WEKA RESULT TABLE ON ADULT DATASET

Parameters	**Naive Bayes**	**J48**	**KNN**	**SVM**
Correctly classified	83.492%	85.9306%	75.9183%	84.9447%
Incorrectly classified	16.508%	14.0694%	24.0817%	15.0553%
Time taken	0.18sec	5.62sec	0.04sec	1275.82sec
Recall	83.5%	85.9%	75.9%	84.9%
Precision	82.6%	85.4%	57.6%	84.2%
F-measure	82.5%	85.5%	65.5%	84.3%

Table4.1: Result of WEKA on adult dataset

According to the above table , KNN take less time for classification as compared to other algorithms and SVM take highest value in time in this classification. The accuracy in classification is higher in J48 and SVM also give good accuracy result as compared to others.

The KNN give the highest incorrectly classified result as compared to others. According to the above table J48 is better as compared to SVM.

Parameters	**Naive Bayes**	**J48**	**KNN**	**SVM**
Correctly classified	82.6211%	89.74%	64.1026%	88.604%
Incorrectly classified	17.3789%	10.25%	35.8974%	11.396%
Time taken	0.01sec	0.09sec	0.02sec	0.038%
Recall	82.6%	89.7%	64.1%	88.6%
Precision	84.2%	89.7%	41.1%	89.1%
F-measure	82.9%	89.6%	50.1%	88.3%

Table4.2: Result of WEKA on ionosphere dataset

According to the above table, Naive Bayes take less time for classification as compared to other algorithms and SVM take highest time for classification. The accuracy in classification is highest in J48 and SVM as compared to others.

RESULTS ON ORANGE WITH ADULT DATASET

Parameters	Naive Bayes	KNN	SVM	J48
Classification Accuracy	88.31%	85.47%	88.32%	88.61%
Precision	88.98%	82.22%	88.33%	90.75%
Recall	93.33%	98.67%	94.22%	91.56%
Time taken	0.04sec	0.02sec	0.05sec	0.03sec
F-measure	91.11%	89.70%	91.18%	91.15%
Sensitivity	93.33%	98.67%	94.22%	91.56%

Table4.3: Result of Orange on ionosphere dataset

According to the above table J48 give the highest accuracy than others. KNN and SVM both are having second highest accuracy. KNN take the less time as compared to others. The precision value is highest in J48 . According to the above result J48 give efficient result in classification on ionosphere dataset .

Parameters	Naive Bayes	KNN	SVM	J48
Classification Accuracy	84.27%	80.06%	No result	82.58%
Precision	71.56%	59.08%	No result	64.72%
Recall	57.56%	55.94%	No result	60.82%
Time taken	16sec	254.8sec	No result	36sec
F-measure	63.80%	57.46%	No result	62.71%
Sensitivity	57.56%	55.94%	No result	60.82%

Table4.4: Result of Orange on adult dataset

According to the above table Naive Bayes give the highest accuracy than other algorithms. SVM don't give any result on this dataset. The time taken is less in Naive Bayes as compared to others.
For this dataset Naive Bayes give best result in classification.

RESULT OF MATLAB ON IONOSPHERE DATASET

Parameters	Naive Bayes	KNN	SVM	J48
Accuracy	82.1%	89.7%	92.0%	90.9%
Execution time	3.87sec	2.581sec	9.98sec	5.56sec
Incorrectly classified	17.9%	10.3%	08.0%	9.1%
Observation speed	2300ob/sec	2000 ob/sec	1500 ob/sec	2100 ob/sec

Table4.5: Result of MATLAB on ionosphere dataset

According to the above table SVM gives the highest accuracy as compared to others algorithms. J48 also give good classification result . The KNN take less time as compared to SVM, Naive Bayes and J48. The number of observation is highest in Naive Bayes.

Parameters	Naive Bayes	KNN	SVM	J48
Accuracy	No result	No result	69.1%	70.5%
Execution time	No result	No result	818.26sec	15.607sec
Incorrectly classified	No result	No result	40.9%	29.5%
Observation speed	No result	No result	5200 ob/sec	9700 ob/sec

Table4.6: Result of MATLAB on adult dataset

According to the above table J48 give highest accuracy as compared to other algorithms. The Naive Bayes and KNN give no result on this dataset. The execution time is less in J48 as compared to SVM. The number of observation is high in J48 than SVM.

COMPARISON TABLE OF TOOLS BASED ON ALGORITHM WITH ADULT DATASET.

Tool(Naive Bayes Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	**83.492%**	**16.508%**	**0.18sec**
MATLAB	**NO RESULT**	**NO RESULT**	**NO RESULT**
ORANGE	**84.27%**	**15.73%**	**16sec**

Table4.7: Result of tools on adult dataset with Naive Bayes algorithm

Tool(SVM Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	84.9447%	15.0553%	1275.82sec
MATLAB	69.1%	40.9%	818.26sec
ORANGE	No result	No result	No result

Table4.8: Result of tools on adult dataset with SVM algorithm

Tool(SVM Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	75.9183%	24.0817%	0.04sec
MATLAB	No result	No result	No result
ORANGE	80.06%	19.94%	254.8sec

Table4.9: Result of tools on adult dataset with KNN algorithm

Tool(J48 Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	85.9306%	14.0694%	5.62sec
MATLAB	70.5%	29.5%	15.607sec
ORANGE	82.58%	17.42%	36sec

Table4.10: Result of tools on adult dataset with J48 algorithm

COMPARISON TABLE OF TOOLS BASED ON ALGORITHM WITH IONOSPHERE DATASET.

Tool(Naive Bayes Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	82.6211%	17.3789%	0.01sec
MATLAB	82.1%	17.9%	3.57sec
ORANGE	88.31%	11.69%	0.04sec

Table4.11: Result of tools on adult dataset with Naive Bayes algorithm

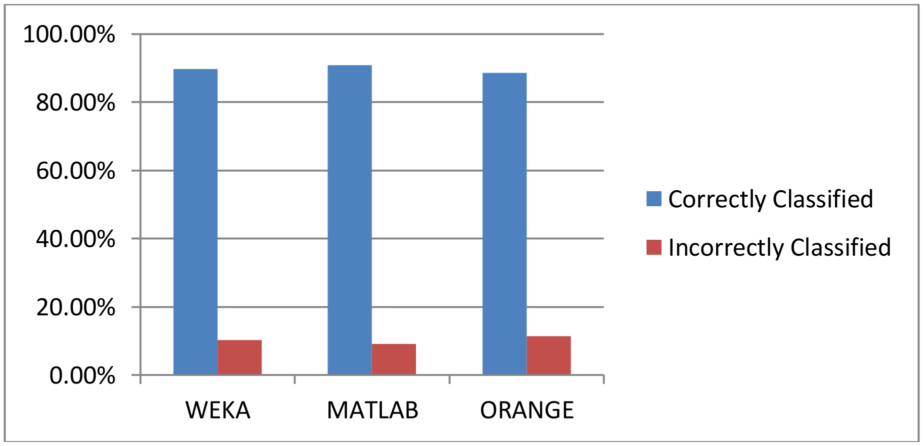


Figure 4.1:Result of tools on adult dataset with Naive Bayes algorithm

Tool(SVM Algorithm)	Correctly Classified	Incorrectly Classified	Time
WEKA	88.604%	11.396%	0.38sec
MATLAB	92.0%	8.0%	9.98sec
ORANGE	88.32%	11.68%	0.05sec

Table4.12: Result of tools on adult dataset with SVM algorithm

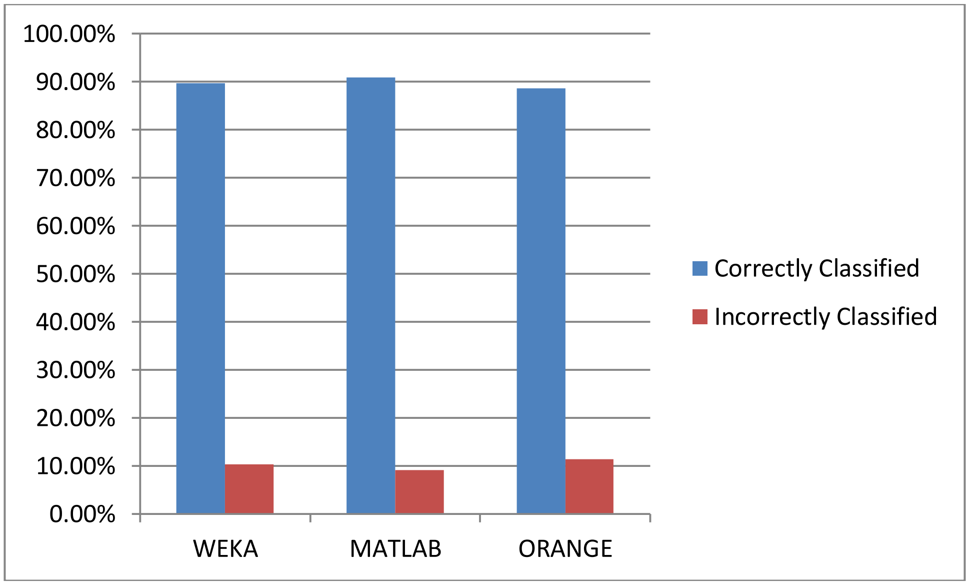


Figure 4.2:Result of tools on adult dataset with SVM algorithm

Tool(KNN Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	64.1026%	35.8974%	0.02sec
MATLAB	89.7%	10.3%	2.581sec
ORANGE	85.47%	14.53%	0.02sec

Table4.13: Result of tools on adult dataset with KNN algorithm

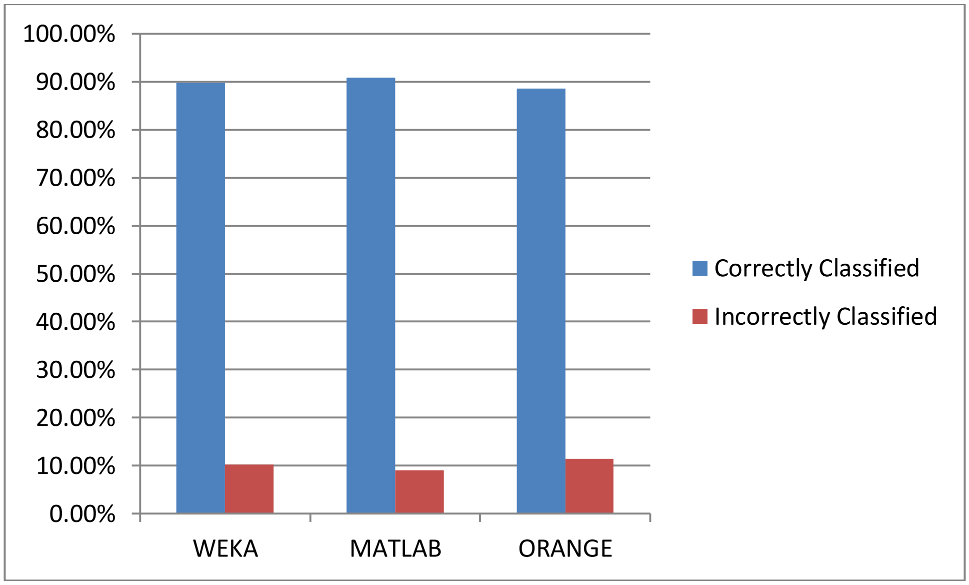


Figure4.3:Result of tools on adult dataset with KNN algorithm

Tool(J48 Algorithm)	**Correctly Classified**	**Incorrectly Classified**	**Time**
WEKA	89.74%	10.256%	0.09sec
MATLAB	90.9%	9.1%	5.56sec

ORANGE	88.61%	11.39%	0.03sec

Table4.14: Result of tools on adult dataset with J48 algorithm

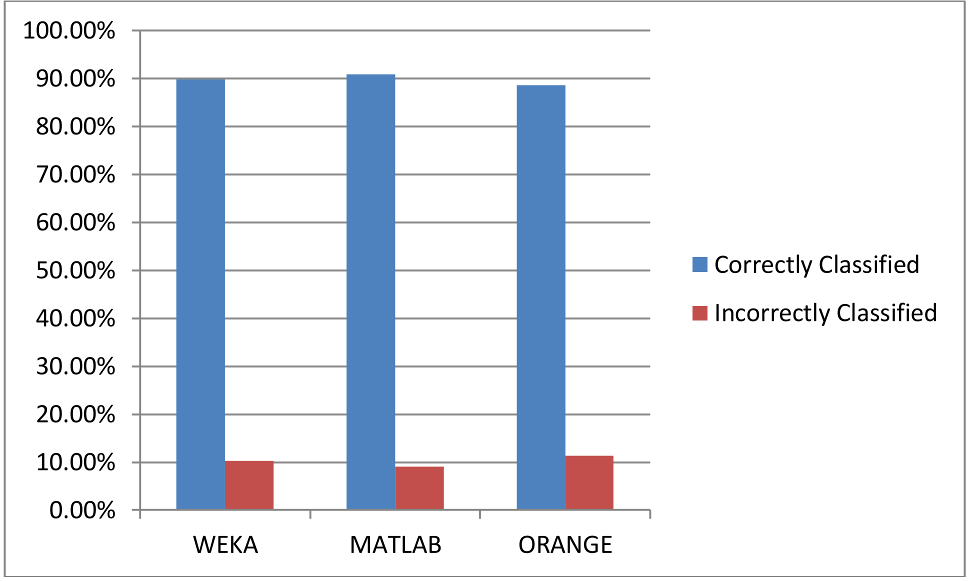


Figure 4.4: Result of tools on adult dataset with J48 algorithm

According to the results on adult dataset with missing value, WEKA give better results than orange and Matlab in accuracy with SVM and J48. On the other side orange give better accuracy than Matlab and WEKA in accuracy with KNN and Naive Bayes. The time taken is less in WEKA as compared to others.

In second dataset i.e. ionosphere the Matlab give better result in SVM,KNN and J48 than WEKA and Orange but the time taken is less in WEKA. The accuracy result is best in Matlab with this dataset.

5.1 CONCLUSION AND FUTURE WORK

Data mining is the process of finding patterns from large amount of data. Most of the companies use the data mining techniques and tools for their benefits. This is used for analyzing and finding information. The data mining tools provide the automatic interface for users to analyze the data. These tools provide many data mining techniques. The major issue in data mining tools is in context of performance. The handling of data and to provide accurate results are two main requirements of good data mining tools. In the thesis, we describe about data mining and discussed about the performance of some of the data mining tools. In the thesis we provided the comparative analysis of three tools namely WEKA, Orange and Matlab on the basis three parameters (correctly classified accuracy, in-correctly classified and time) by applying four algorithms namely SVM, KNN, Decision Tree and Naive Bayes. From these results we have found the variations in accuracy of result and time taken by each data mining tool with respect to each dataset.

In future work we propose the inclusion of new data mining tools along with few more parameters. Testing can be done on the same parameters as well as on new parameters by applying the same algorithms or more algorithms can also be explored. This will be really interesting to find the changing patterns with new additions.

REFERENCES

[1] Jiawei Han, Micheline Kamber Jian Pei. Data mining concepts and techniques.3rd edn. Morgan Kaufmann Elsevier: USA , 2012.

[2] Ian H.Witten, Eike Frank, Mark A.Hall. Data Mining practiced machine learning tools and techniques. 3rd edn. Morgan Kaufmann Elsevier: USA,2011.

[3] "12 data mining tools and techniques", https://www.invensis.net/blog/data-processing/12-data-mining-tools-techniques, 18 November 2015.

[4] "OLAP Tools ",http://www.informationbuilders.com/olap-online-analytical-processing-tools

[5] "10 most popular analytic tools in business", http://analyticstraining.com/2011/10-most-popular-analytic-tools-in-business/,15 January, 2011.

[6] "Defining dashboards, visual analysis tools and other data presentation media",http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/what-is-a-dashboard.aspx,28 November 2011.

[7] "Enterprise Dashboard Digest", http://enterprise-dashboard.com

[8] "Building and Using Dashboards", https://docs.oracle.com/cd/E28 280_01/bi.1111 /e10544/dashboards.htm#BIEUG682.

[9] "What is Apache Mahout", https://mahout.apache.org/

[10] "Teacher Dashboard", http://www.teacherdashboard365.com/

[11] "Orange: Data mining Fruitful and Fun",http://orange.biolab.si/

[12] "Natural language Toolkit",http://www.nltk.org/

[13] "Voyant ",http://voyant-tools.org/

[14] "Alchemy API Tools", http://www.alchemyapi.com/developers/tools

[15] "Alchemy Language",https://www.ibm.com/watson/developercloud/alchemy-language.html

[16] Stavrianou A, Andritsos P, Nicoloyannis N. Overview and Semantic Issues of Text Mining. SIGMOD Record.2007 September

[17] "Introduction to Natural Language Processing" ,http://blog.algorithmia.com/ introduction -natural-language-processing-nlp/,11 August 2016

[18] "Predictive Analytics",http://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/

[19] Markus Hegland," Data mining Techniques", Research gate,cambridge university ,2001.

[20] Amit gupta, Ritul Kumar and Manish Mittal,"Data mining and its application for knowledge mangement", International journal of electrical and electronic engineering and telecommunication, 2015

[21] Kendra J.Ahmed, Mahbub K. Ahmed, Scott Mckay," A brief review of alternative uses of data mining: Education , Engineering and others", ASEEE Gulf-Southwest Annual conference, 2015.

[22] Jigno Ashish Patel ," Classification algorithm and comparison in data mining", IJIACS,2015

[23] Hall M,Frank E , Holmes G, Reutemann B , Witten IH,"The WEKA Data Mining Software: An Update", SIGKDD Explorations.2009.

[24] Demšar J and Zupan B,"Orange: Data Mining Fruitful and Fun - A Historical Perspective",2012.

[25] Berthold M, Cebron N,Dill F, Gabriel T,Kotter T, Meinl T, Ohl P, Sieb C, Thiel K and Wiswedel B,"KNIME: The Konstanz Information Miner",Springer.2008.

[26] Loper E and Bird S ,"NLTK: The Natural Language Toolkit",2002

[27] Haofeng Z ,"ARMiner:A Data Mining Tool Based on Association Rules", Springer,2001.

[28] Kusiak A," Rough set theory: A data mining tool for semiconductor manufacturing ",JANUARY,2001.

[29] Alcalá-Fdez J ,"KEEL: a software tool to assess evolutionary algorithms for data mining problems",Springer,2008.

[30] Christa S, Madhuri K, Suma V,"A Comparative Analysis of Data Mining Tools in Agent Based Systems",2010

[31] Smith G ,Whitehead J, Mateas M ,"Tanagra: A Mixed-Initiative Level Design Tool",ACM ,2010

[32] Mikut R and Reischl M,"Data mining tools", Research gate,2011.

[33] Shelly ,"Performance Analysis of various data mining classification Technique on healthcare data",2011.

[34] Wahbeh A," A Comparison Study between Data Mining Tools over some Classification Methods",International Journal of Artificial Intelligence ,2012

[35] Jain D ,"A Comparison of Data Mining Tools using the implementation of C4.5 Algorithm",International Journal of Science and Research Vol3,2014.

[36] Salma ,"Rule based complaint detection using Rapid Miner",RCOMM; 2013,Volume: 141 - 149,2013.

[37] Arun R and Tamilselvi J,"Data Quality and the Performance of the Data Mining Tool",2015.

[38] Odan H, Daraiseh A,"Open source Data Mining Tools",IEEE.2015.

[39] Shah C, Jivani A,"Comparison of data mining classification algorithms for breast cancer prediction",4th ICCCNT ,IEEE.,2013.

[40] Kakkar P,Parashar A," Comparison of different clustering Algorithm using WEKA tool",International Journal of Advanced Research in Technology, Engineering and Science,2014.

[41] Bavisi S, Ȧ J and Lopes L,"A Comparative Study of Different Data Mining Algorithms",International Journal of Current Engineering and Technology,2014

[42] Gonc P¸ Jr. A,Barros R and Vieira D,"On the use of data mining tools for Data preparation in classification problems",ACIS 11th International Conference on computer and information science ,IEEE ,2012.

[43] Chauhan N and Gautam N,"Parametric comparison of data mining tools" ,IJATES.2015.

[44] Gupta A, Chetty N , Shukla S,"A classification method to classify High Dimensional data",IEEE.2015.

[45] Hassan M , Shahab ME , Hamed EMR,"A comparative study of classification algorithm in E-health Environment",IEEE.2016.

[46] Singh S, Liu Y, Ding W and Li Z,"Evaluation of data mining tools for Telecommunication Monitoring Data using design of experiment",IEEE .2016.